# BioSimCSE: BioMedical Sentence Embeddings using Contrastive learning

**Kamal Raj Kanakarajan, Bhuvana Kundumani, Abhijith Neil Abraham**
**Malaikannan Sankarasubbu**
SAAMA AI Research Lab, Chennai, India
{kamal.raj, bhuvana.kundumani, abhigith.abraham, malaikannan.sankarasubbu}@saama.com

## Abstract

Sentence embeddings in the form of fixed-size vectors that capture the information in the sentence as well as the context are critical components of Natural Language Processing systems. With transformer model based sentence encoders outperforming the other sentence embedding methods in the general domain, we explore the transformer based architectures to generate dense sentence embeddings in the biomedical domain. In this work, we present BioSimCSE, where we train sentence embeddings with domain specific transformer based models with biomedical texts. We assess our model's performance with zero-shot and fine-tuned settings on Semantic Textual Similarity (STS) and Recognizing Question Entailment (RQE) tasks. Our BioSimCSE model using BioLinkBERT achieves state of the art (SOTA) performance on both tasks.

## 1 Introduction

Word embeddings or vector representations of words generated by neural network architectures, capture the semantic relationships between words much better than traditional methods such as one hot encoding, bag of words, and so on. When dealing with large texts in real-world situations, it is essential to capture the semantic relationship between words as well as between sentences. Thus, rich sentence embeddings that capture the overall sentence semantics are critical for NLP systems. Sentence embeddings play a significant role in various NLP tasks such as information retrieval, semantic search, intent detection, natural language inference tasks.

In recent years, pre-trained models with transformer architecture for the general domain have grown in popularity. The advent of BERT Devlin et al. (2018) based models has made generating high-quality vector representations for natural language text much more manageable. These embeddings act as feature inputs for several downstream tasks. However, these models only generate word-level embeddings, from which we can derive sentence-level embeddings by averaging over the word-level embeddings. Another method is to use a cross encoder network from BERT to directly fine-tune for the task.Although this approach outperforms the averaging approach, it is computationally expensive and unsuitable for semantic similarity search and clustering tasks.

The biomedical domain with its corpora, significantly different from the general domain corpora, needs sophisticated and domain-specific models for effective knowledge representation. In this paper, we adapt SimCSE Gao et al. (2021), a state-of-the-art contrastive learning-based sentence embedding method, and release BioSimCSE - a biomedical domain-specific sentence embedding model.

In summary, our contributions are

1. We train and release[1] biomedical sentence embeddings with supervised and unsupervised training objectives from SimCSE.

2. We evaluate our models on biomedical STS and RQE tasks and demonstrate that our BioSimCSE achieves outstanding outcomes in both zero-shot and fine-tuned settings.

## 2 Background

Transformer-based language representations produced by Universal Sentence Encoder Cer et al. (2018) and BERT has aided NLP practitioners and researchers in various NLP tasks. Using BERT, sentence embeddings can either be generated by averaging the context embeddings of the last few layers or from the output of the last layer. SBERT Reimers and Gurevych (2019) shows that sentence embeddings produced by averaging word-level embeddings from BERT-like transformer models are unsuitable for standard similarity measurements

[1] https://github.com/kamalkraj/BioSimCSE



*Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81 - 86
December 7, 2022 ©2022 Association for Computational Linguistics

such as cosine similarity. SBERT uses the siamese network Schroff et al. (2015), a modified BERT network for the generation of fixed-size sentence embeddings. Though SBERT significantly reduces the time during inference and produces quality sentence embeddings, it follows a supervised approach. It heavily relies on labelled data to train sentence embedding models that might not be suitable for domains without labelled corpora.

Natural Language Inference (NLI) datasets are commonly used for supervised training of sentence embeddings models. The Multi-Genre Natural Language Inference (MultiNLI) Williams et al. (2018) corpus is mainly used to train general domain sentence embeddings. MultiNLI has 433k sentence pairs that have textual entailment information annotated. In the biomedical domain, large corpora of text are publicly available as research papers and articles. However, the availability of annotated datasets is lower than that of the general domain, and the number of samples is also low. Medical Natural Language Inference (MedNLI) Romanov and Shivade (2018) and Radiology Natural Language Inference (RadNLI) Miura et al. (2021a) Miura et al. (2021b) are biomedical NLI datasets; merging both yields only 15K sentence pairs for supervised training.

Recent research has explored different training objectives to derive sentence embeddings in an unsupervised manner. Before widespread adoption of transfomer-based models, Skipthought vectors Kiros et al. (2015) and Quick thoughts Logeswaran and Lee (2018) use unsupervised learning to derive sentence representations from unlabeled data with encoder-decoder and encoder architectures respectively. Semantic Re-tuning with Contrastive Tension (CT) Carlsson et al. (2021), BERT-flow Li et al. (2020), Transformer-based Sequential Denoising Auto-Encoder (TSDAE) Wang et al. (2021) and Simple Contrastive Learning of Sentence Embeddings (SimCSE) Gao et al. (2021) propose methods to generate sentence embeddings using a unsupervised approach with different training objectives. A domain like biomedical, where supervised datasets are unavailable, has to rely on unsupervised techniques. In this work, we selected SimCSE because its unsupervised training is comparable to that of its supervised competitors for training sentence embeddings; in addition, SimCSE performed better in our initial experiment with the other unsupervised training objectives described above.

## 3 Methods

The training objective for SimCSE Gao et al. (2021) utilises the contrastive learning approach, which has a cross-entropy loss function with in-batch negatives. In Unsupervised learning, positive pairs are made by giving the same sentence to the pre-trained encoder twice with regular dropout as noise, all other sentences in a mini-batch act as negative pairs. The NLI dataset has a contradiction hypothesis for each premise and its entailment hypothesis. For supervised sentence embeddings training, SimCSE uses entailment pairs as positive cases and adds matching contradiction pairs and other sentences in the mini-batch as negatives. As in the original SimCSE implementation, we use the `[CLS]` (first token of the sequence) as sentence embedding. Unsupervised SimCSE uses `[CLS]` with an MLP layer (only used in training), and supervised SimCSE uses `[CLS]` with MLP.

We initialize our sentence embeddings models from state-of-the-art transformer model, PubMedBERT Gu et al. (2020) and BioLinkBERT Yasunaga et al. (2022) from Biomedical Language Understanding and Reasoning Benchmark (BLURB) Gu et al. (2020) for our experiments. We excluded ELECTRA Clark et al. (2020) variants BioELECTRA Kanakarajan et al. (2021) and BioM Alrowili and Shanker (2021) from the BLURB because the quality of embeddings created by ELECTRA due to its Replaced Token Detection pre-training task is poor as shown in COCO-LM Meng et al. (2021).

Using biomedical corpora detailed in 3.1, we train biomedical domain-specific unsupervised and supervised sentence transformer models. The sentence embeddings training is done only with the model base architecture - 12 layers of transformers block with a hidden dimension of 768 and multi-head attention over 12 layers. Hyper-parameters used for training are provided in Appendix A. The trained sentence embeddings are then evaluated in zero-shot and fine-tuned settings on the three datasets outlined in 4.2.

### 3.1 Training data

We obtained 1 million sentences randomly sampled from PubMed Central (PMC) [2] published as of April 2022. Pubmed Parser Achakulvisut et al. (2020) is used to extract the abstracts and SciSpacy Neumann et al. (2019) for sentence tokenization. This data is used for unsupervised model training.

[2] https://www.ncbi.nlm.nih.gov/pmc/

The MedNLI dataset comprises sentence pairs annotated for contradiction, neutrality, and entailment by physicians from the Past Medical History section of MIMIC-III Johnson et al. (2016) clinical notes. The dataset contains 11,232 training, 1,395 validation, and 1,422 test cases. The RadNLI dataset contains annotated sentence pairings from the MIMIC-CXR database Johnson et al. (2019). The dataset includes a validation set of 480 and a test set of 480 pairings. For supervised model training, we merge the training, validation, and test sets from these two datasets.

## 4 Evaluation

We use STS and RQE tasks in the biomedical domain to evaluate the performance of our BioSimCSE sentence embeddings model. The datasets used for evaluation are detailed in 4.2. The similarity between two sentence embeddings is determined using cosine similarity. We determine a threshold in cosine similarity using the development set to classify entailment or not for RQE (binary classification) dataset. Using Spearman's correlation, we evaluate STS in line with the original SimCSE research. For RQE, accuracy is used. We evaluate BioSimCSE sentence embeddings under zero-shot and fine-tuned settings.

### 4.1 Evaluation Settings

In a zero-shot setting, the trained supervised and unsupervised BioSimCSE model is used to derive the sentence embeddings directly and evaluate the tasks. In the fine-tuned setting, With task-specific datasets, we further fine-tune the supervised and unsupervised trained BioSimCSE models to adapt better to the task's requirements for the sentence embeddings. For fine-tuning, the sentence embeddings $(u, v)$ for each pair of sentences are derived. Using mean squared loss as the objective function for STS datasets and contrastive loss for question entailment datasets, we optimize cosine similarity between $(u, v)$. Hyper-parameters used for fine-tuning are provided in Appendix A. The fine-tuned sentence embeddings are evaluated only with the corresponding task used for fine-tuning.

The results for zero-shot and fine-tuned are shown in table 1. We also train cross encoder, in which the transformer model takes two sentences and predicts a similarity score or a classification label, as described in the BERT Devlin et al. (2018) paper. This is the standard approach for STS and RQE (Binary classification) tasks. Results for cross encoder is available in table 2. We only compare our models to biomedical-specific models because recent research Gu et al. (2020) has shown that models pretrained with biomedical domain-specific corpora perform significantly better than general-domain language models for Biomedical NLP tasks.

### 4.2 Evaluation Data

BIOSSES dataset provides a collection of 100 similar sentence pairs manually annotated in the biomedical domain. We use the train-test split from BLURB Gu et al. (2020), 64 pairs for training, 16 pairs for validation and the remaining 20 pairs for testing. ClinicalSTS is the STS task in the clinical domain, the latest version provided by n2c2 2019 challenge Wang et al. (2020) has 1641 samples for training and a test set of 412 samples. We use the test set for evaluation, and we have split 1641 samples into 80% train and 20% validation set. Finding entailment between two questions in the context of QA is the objective of RQE. 8,588 training pairs and 302 testing pairs in the initial release Abacha and Demner-Fushman (2016). We use the MEDIQA 2019 Challenge Ben Abacha et al. (2019) test set as the testing pair and the original as the development set.

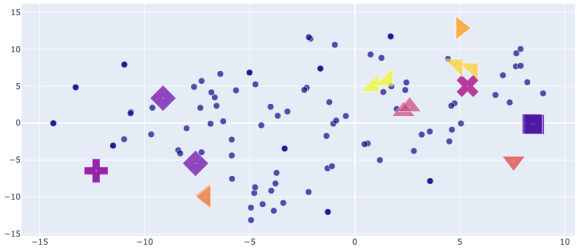

Figure 1: The t-SNE of sentence representation of BioLinkBERT before training with SimCSE

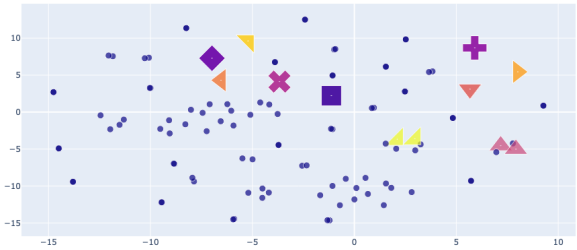

Figure 2: : The t-SNE of sentence representations after with training Unsupervised SimCSE. Similar pairs are denoted by identical shapes. The points are drawn from ClinicalSTS's most semantically comparable sentence pairings (with 5-score labels).

| | | Zero shot | | | fine-tuned | | |
|---|---|---|---|---|---|---|---|
| | | **BIOSSES** | **ClinicalSTS** | **RQE** | **BIOSSES** | **ClinicalSTS** | **RQE** |
| Sent2vec | BioSentVec | 77.98 | 48.72 | 51.56 | - | - | - |
| BioSimCSE$_{\text{Supervised}}$ | PubMedBERT | 83.13 | 72.17 | 53.04 | 85.91 | 77.87 | 56.52 |
| | BioLinkBERT | 90.32 | 76.42 | 54.35 | 92.73 | 81.35 | 57.39 |
| BioSimCSE$_{\text{Unsupervised}}$ | PubMedBERT | 90.61 | 80.67 | 51.94 | 93.57 | 81.16 | 56.52 |
| | BioLinkBERT | 94.55 | 81.02 | 56.61 | **96.37** | **83.76** | **60.04** |

Table 1: Results on BIOSSES, ClinicalSTS and RQE test sets as described in 4. Metric, Spearman's correlation for BIOSSES and ClinicalSTS and accuracy for RQE.

| | **BIOSSES** | **ClinicalSTS** | **RQE** |
|---|---|---|---|
| PubMedBERT | 89.94 | 79.28 | 51.73 |
| BioLinkBERT | 91.75* | 80.42 | 53.47 |

Table 2: Results on cross encoder architecture. * Current state of the art (SOTA). Metric, Spearman's correlation for BIOSSES and ClinicalSTS and accuracy for RQE.

## 5 Results

The BioSimCSE$_{\text{unsupervised}}$ BioLinkBERT model achieves remarkable results on all three datasets during the zero-shot evaluation. The zero-shot performs even better than BioLinkBERT fine-tuned with task-specific data using cross-encoder architecture. From the t-sne of sentence representation Figure 2, we can see that the similar sentence pairs (denoted by the same shapes) are closely aligned after training the BioLinkBERT model with SimCSE and also the average cosine similarity increased from 86.5 to 90.1 for the same. We can also observe that the transformer-based models have outperformed BioSentVec Chen et al. (2019), a non-transformer-based model with a large margin for both BIOSSES and ClinicalSTS. BioSentVec utilizes word vectors and n-grams approach to generate sentence embeddings using sent2vec Pagliardini et al. (2018) model. From the results, we can see that the supervised training of SimCSE is not practical as the unsupervised training, as the biomedical domain has a limited no.of samples in the NLI dataset.

When fine-tuned with task-specific data BioSimCSE$_{\text{unsupervised}}$ BioLinkBERT model sets new state-of-the-art results for all three datasets. For BIOSSES, Spearman's correlation is improved by +4.62 points, compared to the previous SOTA of 91.75. For ClinicalSTS the current SOTA is by BioELECTRA Kanakarajan et al. (2021) 82.11, BioSimCSE improve the SOTA by +1.65 points. BioSimCSE improve the RQE baseline 54.1 accuracy score Abacha et al. (2019) by +5.94 points and sets a new SOTA. We have omitted the RQE SOTA result from PANLP Zhu et al. (2019) (accuracy of 74.9), as this score is achieved using multitask and ensemble methods.

Performance on the evaluation datasets has steadily improved for both BioLinkBERT and PubMedBERT following training with SimCSE compared to the cross-encoder approach.

## 6 Conclusion

In our work, we have explored SimCSE for training sentence embeddings in the biomedical domain. We utilize the publicly available biomedical literature and NLI dataset for training the network in an unsupervised and supervised fashion and further fine-tune them with the task-specific datasets to adapt better to the task's requirements. Our BioSimCSE model has achieved SOTA on all three evaluation datasets. Our results demonstrate that SimCSE unsupervised training objectives can be able to train high-quality biomedical domain-specific sentence embeddings. We make the code and weights available for all of our models for reproducibility.

## Limitations

In our experiments, we have only considered transformer base size models, whereas the Original SimCSE work evaluated both base and large size models. The sample sizes of the datasets used to evaluate sentence embeddings are limited. However, these are the biomedical domain's only sentence pair datasets. After training with SimCSE, the models have only been evaluated on sentence pair similarity/classification tasks and not on any classification of single sentence tasks.

## References


Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2016:310–318.

Asma Ben Abacha, Chaitanya P. Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *BioNLP@ACL*.

Titipat Achakulvisut, Daniel Acuna, and Konrad Kording. 2020. Pubmed parser: A python parser for pubmed open-access xml subset and medline xml dataset xml dataset. *Journal of Open Source Software*, 5(46):1979.

Sultan Alrowili and Vijay Shanker. 2021. BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021a. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021b. Radnli: A natural language inference dataset for the radiology domain.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

*Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning.

Y. Wang, S. Fu, F. Shen, S. Henry, O. Uzuner, and H. Liu. 2020. The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview. *JMIR Med Inform*, 8(11):e23375.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

Wei Zhu, Xiaofeng Zhou, Keqiang Wang, Xun Luo, Xiepeng Li, Yuan Ni, and Guo Tong Xie. 2019. Panlp at mediqa 2019: Pre-trained language models, transfer learning and knowledge distillation. In *BioNLP@ACL*.

## A Example Appendix

The learning rate and batch size for training SimCSE supervised, and unsupervised models are the same as the original work. Our search for hyperparameters also shows that these give the best results. Both supervised, and unsupervised training was done using 512 batch sizes and learning rates 1e-5 and 5e-5, respectively. For the unsupervised model, we train the model with 1 million and 2 million examples, and we use zero-shot sentence similarity to measure how well it does. For one epoch, the model was trained. Adding more data after 1 million does not make a big difference in performance compared to the cost of training the model. The sequence length is restricted to 128 tokens in all our experiments. We use the SimCSE[3] implementation that the authors made available as open source to train our sentence embeddings. All the experiments are done using a single NVIDIA Titan RTX (24GB VRAM) GPU.

Table 3 lists all of the hyperparameters used for task specific fine-tuning of sentence embedding and cross encoder fine-tuning.

| | Hyperparameters |
|---|---|
| Epochs | 3-20 |
| Learning rate | 1e-5, 2e-5, 5e-5 |
| Batch size | 8, 16 |

Table 3: Sentence embeddings and cross encoder fine-tuning hyperparameters

Figure 3 shows how the similarity of sentence pairs is computed using the cosine similarity metric. The standard cross encoder architecture used with transformer models for sentence pair tasks is shown in Figure 4.

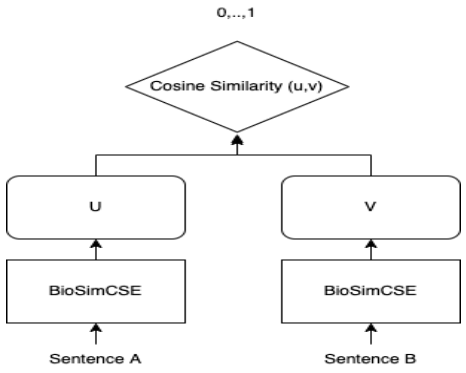


Figure 3: Finding similarity of sentence pair using BioSimCSE model

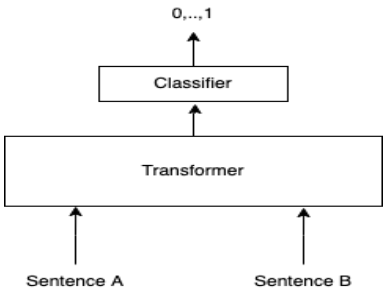


Figure 4: Cross encoder fine-tuning for sentence pair regression/classification

[3]https://github.com/princeton-nlp/SimCSE